Generating /p/,/t/,/k/ consonants by a physical modeling of musical percussion

Patrick Fourcade, Claude Cadoz

Generating /p/, /t/, /k/ consonants by a physical modelling of musical percussion.

Patrick Fourcade and Claude Cadoz, ACROE, 46, av. Felix Viallet, France 38031 Grenoble cedex.
Email: patrick.fourcade@imag.fr, claude.cadoz@imag.fr

Abstract: We study the percussion mechanism in music instruments and we research some perception and production cues for the paradigm of the instrumental relation for artistic creation. According to ecological assumptions, our investigation relies on a computer language by physical modeling. We make a connection between unvoiced plosive consonants and a model of a musical percussion. We attempt to synthesise /p/, /t/, /k/ consonants by a percussion model. We develope an original vibrating structure for vowel synthesis. Thus, we explore a new category of instrumental excitator : "semipercussive" and semimaintained. A short psychoacoustic evaluation allows to say : maintained excitation for inducing vowels contribute indirectly to the plosives recognition. Otherwise, isolated percussion do not. We conclude with general considerations on the mechanism of prototype categorisation.

1 Introduction

Music has its origin in the human body. It can be separated into vocal music and instrumental music**[Schaeffner 1968a]**. Many instrumental partitions refer to sung voice and sometimes also to spoken voice.

Despite this classical separation, we intend to show a connection between the intimate process of sound production in a physical instrument and the perception of some phonemes. This attempt of connection does not mean that the authors claim to contribute to the field of phonetics. This work is relevant in the following problematic: Research on expressive tools for musical creation.

In the psychological theory of data processing**[Anderson 1985, Lindsay and Norman 1977]**, the recognition of the identity and meaning of a sound event is the result of analysis, fit and association operations. Another approach is the theory of the ecological perception**[Michaels and Carello 1981]**. The human perception consists in extracting cues useful for survival. In this manner, the auditory system has developed some capacities in a world where sounds are practically all have a mechanical origin. According to the previous theory, we use a musical instrumental representation based on dynamics phenomena.

Focusing on musical percussion, we notice in previous investigations that some synthesis percussion sounds have the property of inducing plosive consonants. Searching relevant invariants for perception and production in percussion sounds, we transpose plosive consonants and the human phonatory organ in the field of instrumental sound modelling. Moreover, phonemes compassed of these consonants are interesting for studying because, within the same sound entity, coexist "percussive" excitation and maintained excitation.

Unlike the vowels' steady sound, consonants are more difficult to analyse and to describe acoustically. There are rapid and subtle changes in the sound. We focus on the unvoiced plosive consonants. We show a musical instrument model for the synthesis of phonemes using /p/, /t/ and /k/ consonants.

From an instrumental point of view, the sound is the output of a physical chain achieving a succession of energetic transformations and transfers. A music instrument is made of three parts: the excitator, the vibrator and the resonator. The vibrator is a passive vibrating structure with very low dissipation (strings in the violin, air column in the clarinet or in the vocal tract). The nature of the sound of the instrument is strongly determined by the vibrator. The excitator is the part of the instrument between the human action and the vibrator. It communicates energy to the passive vibrator. We distinguish two classes of excitators: - maintained excitators transmit a continuous energy during time (bow in the violin, vocal cords for speech); - "percussive" excitators transmit a brief energy (key-hammer in the piano, plectrum in the guitar). The resonator (harmony board in the piano, vocal tract for speech, ...) makes an adaptation between the vibrating phenomena in the vibrator and the output environment.

First, we exhibit the environment of synthesis for sound and the methodology used in the research of musical creation. Next, we explain a vibrator model which generates vowel patterns. Third, we explain a percussion excitator model inducing /p/, /t/ and /k/ percepts. Finally, putting together the excitator and the vibrating structure, we present the instrumental representation of phonemes. Then, a psychoacoustical experiment attempts to validate the percussion model.

2 The reasons for a research in physical modelling for musical creation - The CORDIS-ANIMA language .

2.1 Purpose

Since the beginning of sound synthesis, many experiences have shown that relations between acoustical description and auditory perception were not simple**[Castellengo 94]**. Psychoacoustics of complex sounds reveals listening strategies depending on an ecological paradigm**[Risset 1994]**. We make the assumption that instrumental relation, including physical properties, is a relevant paradigm for musical creation. Putting the computer machine and the human-machine interaction to the heart of the matter, since 1975, our laboratory has chosen to use the theoretical power and the formalism power of computer science, especially algorithmics in real time.

Within the framework of the realisation of a computer environment for musical creation, the laboratory have developed a system for sound synthesis, based on the physical modelling of musical instruments. In this perspective we have been driven to study the problem of analysis and the question of interpretation of sound phenomena in relation with the physical modelling principles used for synthesis.

2.2 Language

Physical modelling is interesting because it offers a convenient and convincing vision of sound synthesis for both scientists and musicians. The CORDIS-ANIMA system developed at the ACROE **[Cadoz, et al. 1993]** allows for computer-based modelling and simulation of physical objects that can be seen, heard and handled (with force-feedback gestural control device). An object is a modular assembly of elementary mechanical components picked up amongst a limited number of types with very simple associated elementary simulation algorithms. The description is based on the fundamental laws of Newtonian Mechanics. Their assembly constitutes the global simulation algorithm of the modelled object. A simulation is the computation of all displacements and force transfers inside the object. This general formalism enables a great flexibility for the construction of virtual objects which may or may not have a referent in the real world.

2.3 Methodology

The investigation method is based on three levels of sound representation: the human perception, the signal sound and the physical representation. We analyse the sound phenomena by confronting these three levels.

Natural sounds play the role of metaphors. They do not constitute a proper goal in synthesis. Furthermore, CORDIS-ANIMA is a modelor-simulator. The investigation exploits the relation: simulate to know, know to simulate.

CA is a research axis of our laboratory. Algorithm representation, retrospectively, throws light on sound phenomena, especially when the task consists in a representation economy. Exploring sound possibilities of the CA language for itself is also justified.

2.4 Realisations

CA shows good capacities for the description of virtual sound objects. Some matters such as wood, or metal are evoked by vibrating structures. Some shapes and volumes such as plates, bars, membranes, strings, air columns are evoked from intuitive and simple models**[Cadoz and Incerti 1998, Incerti 1996]**. Some excitator models generate sounds evoking bowed string, violin bow**[Florens and Cadoz 1990]** , wind instruments such as clarinet, gong percussion, xylophone.

The models mentioned above are relevant in accordance with three criteria:

-1- realism of the instrumental effect;

-2- the possibility of playing on the virtual instrument, accessing a wide variety of timbres linked with gestures;

-3- the capacity of generating, from an initial father instrument model, realistic son models without any known references.

Focusing on the category of "percussive" excitator systems, we are going to analyse the specific problem of plosive consonants.

3 Construction of vowels structures

In a physical way, we have to build a vibrating structure inducing vowels. We know that the acoustical structure of vowels is constituted by stationary sounds. Sound vowels can be represented as signals coming from the filtering of an active continuous energy source. The filter is a spectral envelope: the formants.

Vowels taken into account in this study are /a/, /ɛ/, /i/ and /u/. The character of vowels' sound is mainly determined by the first and second formants. So we limit vowel description to the first and second formants. Relative amplitudes and bandwidths of formants are neglected.

For the passive part, we model a vibrating structure with an impulse-like frequency response similar to the formant structure of the vowel's spectra.

We consider a line composed of three masses with one free end moving in a one dimensional space. The line is homogeneous. We adjust the parameters values in order to make the first two spectral modes correspond to the formants. In a simple manner, this line models the vocal tract. For the physicist, it can be considered as a pipe closed at one end (by the glottis) and open at the other end (the lips). To improve the model, each peak of the line spectrum needs thickening. To do that, we use a specific method developed in our laboratory **[Incerti 1997]** carrying out the algebraic "product" of the line with a twenty-mass-agglomerate. We call the result VS (Vibrating Structure) (cf. FIG. 1).

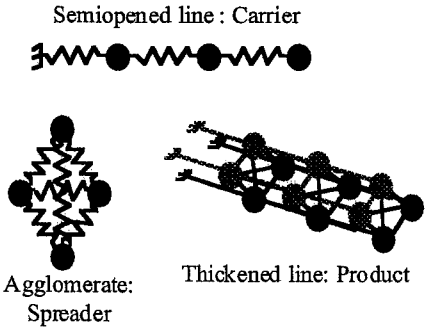

FIG. 1 - *Vibrating structure VS; a thickened line resulting of the product of a semiopen line by an agglomerate (Simplified example - four-masse-agglomerate instead of twenty, viscous elements are not represented).*

Thus, the frequency response of VS is closer to the vowel's acoustical formant structure. The stability of formants is maintained by parcels of modes. Parcels are concentrated around frequency peaks of the line which is tuned by the first two formants of the vowel.

To produce a given vowel, a simple excitation of VS is not satisfactory. It is similar to the sound made by hitting the cheek with a finger, the vocal tract being on a specific vowel configuration. We need a maintained excitator. It will be described in section 4.2.

4 Construction of /p/, /t/, /k/ excitators

Many authors focus their works on the task of consonant recognition and suggest acoustical cues. Using the sonogram representation of phonemes, a reference study **[Cooper, et al. 1952]** showed that /p/, /t/ and /k/ consonants can be generated by time-frequency atoms (bursts) in front of the formant structure of vowels as:

- /t/ consonant can be synthesised by a high frequency pattern;
- /p/ consonant need a low frequency pattern equal to the frequency of the first formant;
- /k/ consonant need an intermediate frequency pattern just a little over the centre frequency of the second formant.

The duration of consonant transients was about 15 ms.

Another experiment with the same synthesis technique demonstrates that the modification of the second formant during the transient phase (continuous trajectories of the formant) enables to synthesise the /p/, /t/ and /k/ consonants without using burst **[Delattre, et al. 1955]**. Anyway, we just exploit the results of the F. Cooper experiment.

4.1 The consonant, a "percussive" attack transient

The percussion model for consonant is considered as a free point mass Mp. All movements are one dimensional. The interaction linking mass Mp and VS is a repulsive conditional force of stiffness Kp. The interaction is such that F = Kp.(Ysv-Yp) if Yp<Ysv, F = 0 otherwise. Yp is the altitude of the mass Mp; Ysv is the altitude of the VS impact mass. We call the percussion model EX. Initially, the EX mass has a velocity and the VS is at rest. The EX model has been studied exhaustively in **[Fourcade and Cadoz 1996]**.

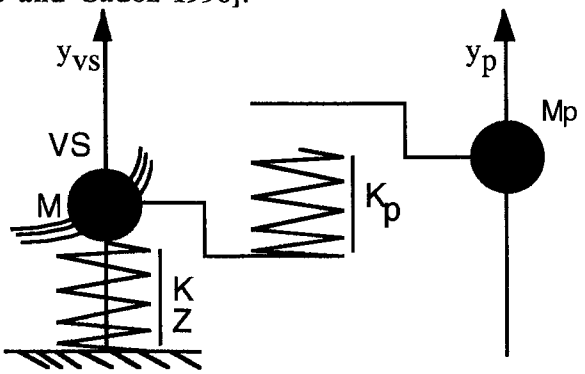

FIG. 2 - *EX model. The inertia Mp strikes one of the VS masses with the interaction law : F = Kp.(Ysv-Yp) if Yp<Ysv, F = 0 otherwise*

To reproduce /p/ consonant, we make the assumption that the behaviour of VS to a first-order oscillator, O1. We consider the EX-VS coupled system: the striker of inertia Mp, the spring of stiffness Kl and the oscillator O1. In this case, in adjusting modal parameters, it is possible to lay down the system **[Mangiarotti 1997]**:

-1- One frequency for the higher mode equals the frequency of the first formant;

-2- One frequency for the lower mode such that the half period equals 15 ms (transient duration proposed in the Cooper experiment).

So we determine the inertia Mp and the stiffness Kl of the striker.

In order to reproduce the /t/ consonant, this technique is not convenient because the previous approximation on VS is not correct. We have tuned in an empirical manner the parametric frequency of the striker fp ($2\pi f_p = \sqrt{k_l / M_p}$) in order to obtain a coupling frequency much higher than the second formant (cf. **[Cooper, et al. 1952]**). The inertia of the excitator is tuned in an empirical manner and leads to a transient of 2 ms.

In order to reproduce the /k/ consonant, we tune the mode parameter of the buffer in order that the coupling frequency will be lightly higher than the second formant (cf. **[Cooper, et al. 1952]**). The /k/ percept appears just only for a weak inertia of the striker. The transient has a duration of

1 ms. Meanwhile, if this inertia is too weak, resulting sound blends with an instantaneous impulse-like; duration of transient tends to zero.

A series of sounds shows the results of consonant synthesis - one percussion on VS without the maintained excitation.

4.2 Maintained excitator

However, we need to tune up a maintained excitation. The process consists of sending, at regular periods of time, a moving inertia that strikes the vibrating structure in a same point according to the buffer law (cf. FIG. 2). The time gap between two percussions is small enough such that the auditory system does not perceive the percussion effect but a continuous process. The gap is 10 ms.

We can notice that this kind of excitation is similar to the application of a "comb" on a vibrating object. The percussion instruments call "racles" (scrapes)**[Schaeffner 1968b]** are the natural reference of this excitation type. On the contrary, for a bow excitator type (the violin bow for example), the excitator is not influenced by the vibrating structure behaviour.

4.3 Making phonemes: Excitation by percussion followed by a maintained excitation

We make precede the generating vowels' maintained excitation by a "percussive" excitation whose parameters enable to induce the transient of unvoiced plosive consonants.

The excitation at the attack communicates to the VS a higher energy than maintained excitation. Even if there is always 10 ms between the first impact and the second one, the nature of the excitation at the attack distinguishes from the maintained excitation. For all simulations, the ratio between the inertia of successive strikers, Ms, and the inertia of the transient excitator, Mp, controls the *attack ratio*. Attack ratio means the relative sound level between attack and the following part of the sound. The inertia Ms is tuned in order to give to the transient a "natural" attack ratio. The frequency parameter of successive percussion excitators, fs, is such that the succession of impacts is perceived as a continuum.

TAB. 1 - *Parameters of the "percussive" excitator and the maintained excitator: Mp and fp, inertia and modal frequency of the first percussion excitator; Ms and fs, inertia and modal frequency of the successive percussion excitator following.*

	[pa]	[pe]	[pi]	[po]
Mp	91,67 Kg	68,47 Kg	36,66 Kg	43,66 Kg
fp	67,62 Hz	67,93 Hz	68,89 Hz	68,57 Hz
Ms	60 Kg	60 Kg	60 Kg	60 Kg
fs	35,59 Hz	35.59 Hz	35.59 Hz	35.59 Hz

	[ta]	[te]	[ti]	[to]
Mp	10 Kg	10 Kg	10 Kg	10 Kg
fp	3000 Hz	2500 Hz	3000 Hz	3000 Hz
Ms	2 Kg	12 Kg	2 Kg	2 Kg
fs	300 Hz	100 Hz	300 Hz	300 Hz

	[ka]	[ke]	[ki]	[ko]
Mp	0,2 Kg	0,2 Kg	2 Kg	2 Kg
fp	1250 Hz	1550 Hz	2000 Hz	650 Hz
Ms	0,05 Kg	0,8 Kg	2 Kg	1 Kg
fs	300 Hz	100 Hz	300 Hz	300 Hz

The results of phonemes synthesis are given by a series of sounds. Sonograms FIG. 3 and FIG. 4 of phonemes synthesis [pa], [ta], [ka] and [pe], [te], [ke] show the connection with the results of Cooper.

The /p/ transients, from sonograms, has a strong energy low partial frequency; excited frequencies are mainly smaller than 2000 Hz. We observe that /k/ and /t/ transients excite formant peaks in the same manner. The main difference is based on: - first, /t/ transient contains

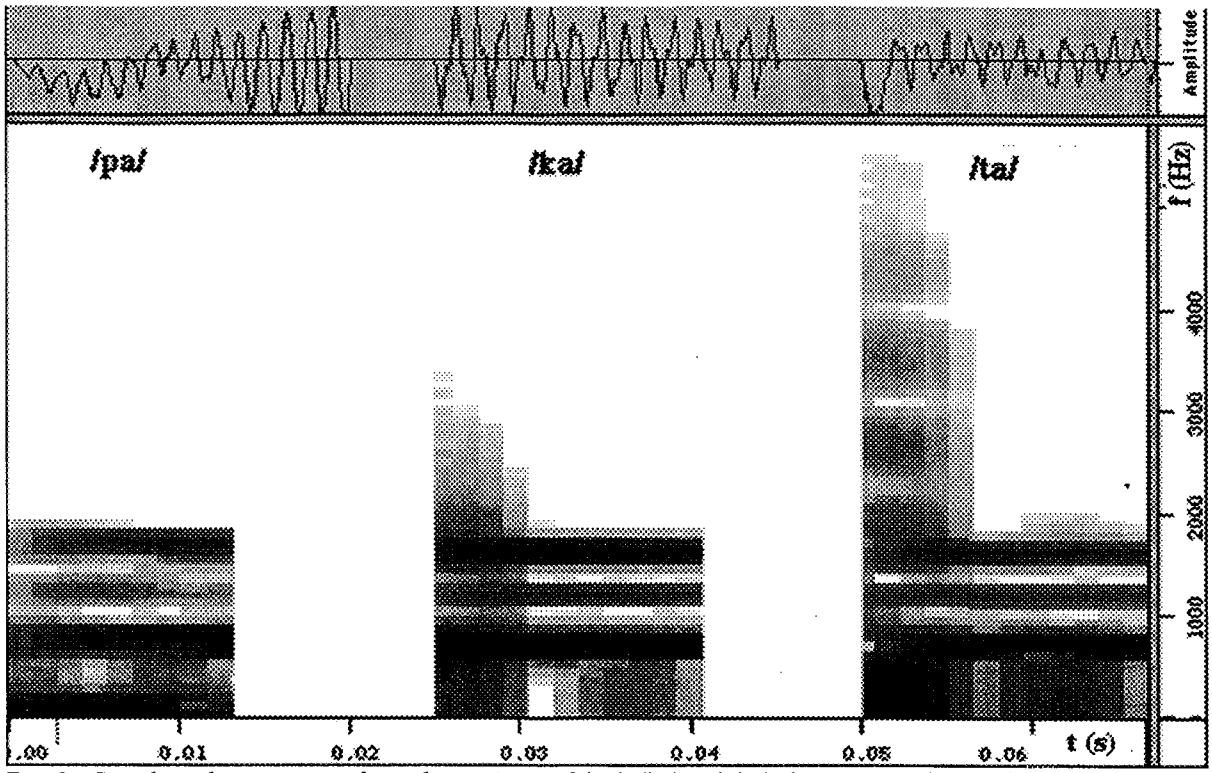

FIG. 3 - *Signals and sonograms of attack transients of /pa/, /ka/ and /ta/ phonemes synthesis. The first transient is different because it has a strong energy at low frequencies. The centre transient has a pattern upon the second formant. On the right, the transient has high frequencies during a very short time.*

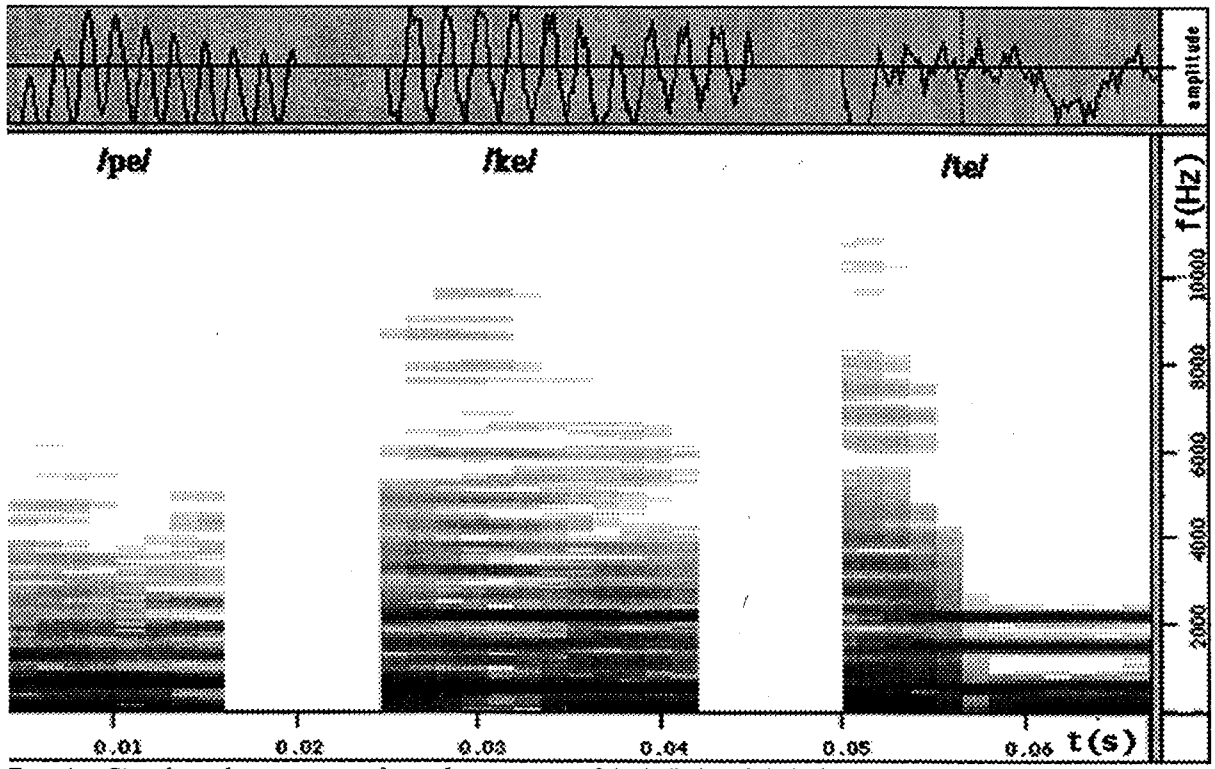

FIG. 4 - *Signals and sonograms of attack transients of /pe/, /ke/ and /te/ phonemes synthesis. The left transient energy is concentrated on the first formant. The middle transient reveals a quite homogeneous excitation. The right transient has a* cgs *which falls down suddenly in the time.*

higher frequencies; - next, the phoneme resonance beginning with /t/ is suddenly impoverished. This is particularly obvious for /te/ and /ke/. Considering the spectral gravity centre of a signal sound (cgs), these differences reinforce the assumption that cgs temporal variation is an important parameter for the perception of instrumental timbres**[Freed 1990]**.

The information on the phase signal is not taken into account because the sonogram represents only the modulus of the Short-Time-Fourier-Transform. The phoneme representation by physical modelling gives an information that plainly describes the temporal process implied.

5 Recognition of synthesis /p/, /t/ and /k/ percepts

A first experimentation consists of presenting 12 percussion stimuli without maintained excitation. Three percussion excitators, according to /p/, /t/ and /k/ parameters, are applied respectively on four passive VS, tuned as the vowels /a/, /ɛ/, /i/ and /o/. For each stimulus, the subject is invited to recognize the synthesised consonant: /p/, /t/ or /k/ and to give a degree of comparison relevance on a scale of 1 to 8. On ten subjects, degree of recognition is under the stochastic rate. The EX model is not relevant for inducing isolated /p/, /t/ and /k/ synthesis.

A second experimentation presents the same percussion physical models but with maintained excitations on VS. The subject is invited to recognise phonemes /pa/, /ta/, /ka/, /pe/, /te/, /kè/, /pi/, /ti/, /ki/, /po/, /to/ and /ko/. We obtain a recognition rate of 7/12 for the consonants, of 8/12 for the vowels and of 3/12 for the phonemes. Stochastic rate is respectively of 4/12, 3/12 and 1/12. We deduce that the presence of maintained excitation allows plosives /p/, /t/ and /k/ recognition. Moreover, the degree of recognition for the phonemes is three time greater to the stochastic rate.

6 Conclusion

Searching, in the vocal expression domain, some "percussive" phonemes, we have elaborated a synthesis model for unvoiced plosive consonants /p/, /t/ and /k/. We simulate corresponding phonemes. Compared with traditional synthesis techniques for speech, the specific interest of our process consists in using a physical representation centred on the dynamic phenomena.

We have been led to build a physical model for generating some vowels. We have shown that the concept of *product structures* could be applied with accuracy to the construction of acoustical structures constituted of formants.

A percussion model allows to generate bursts describing in the Cooper experiment. Thus, unvoiced plosive consonants can be considered, in the instrumental universe, as percussion. Phonemes build from /p/, /t/ and /k/ transients can be considered as the musical play of a virtual instrument: a "thickened" semiopen chaplet excited by a point mass percussion followed by very close series of percussion.

The model is interesting for the musical creation according to three points: - the psychoacoustic study reveals a relative success on the plosives effect; - a consonant human gesture should be realised for music playing; - an intermediate excitator combining "percussive" and maintained excitation constitutes a first step in this domain.

Even if the virtual percussion instrument is in the limited field of plosives, it already gives to musicians an original tool of sound composition. Phonemes production for Tabla players is sometimes integrated as an element of instrumental effect. The percussion model offers to immerse speech production in the instrumental universe, in the deep "privacy" of the vibrating matter.

Furthermore, this work is a contribution to the theoretical notion that there are general principles of perception-production relationships that extend across both speech and non-speech domains**[Gibson 1966]**. From theoretic studies of **[Rosch 1978]**, the mechanism of prototype categorisation is based on two principles:

> -1- "the cognitive economy, which expresses the necessity to access maximum information for a minimum cognitive effort";
>
> -2- "The non-equiprobability of perceived events, which takes into account the organisation of the world".

Does speech, as a communication tool, obey to the principle of cognitive economy ? Does the instrumental representation exploit the principle of non-equiprobability ? Anyway, our conviction is that it will be interesting to carry on this comparison for all consonants: fricative consonants and maintained excitator model, voiced plosive consonants, semivowels, liquids...

References

[Anderson 1985] J. R. Anderson, *Cognitive psychology and its implications*. New York: Freeman, W. H., 1985.

[Cadoz, et al. 1993] C. Cadoz, A. Luciani and J.-L. Florens, "CORDIS-ANIMA : a Modeling and Simulation System for Sound and Image Synthesis - The General Formalism," *CMJ*, vol. 17, 1993.

[Cadoz and Incerti 1998] C. Cadoz and E. Incerti, "Synthèse musicale par modèles physiques : modélisation de structures vibrantes avec le langage CORDIS-ANIMA," in *Recherches et applications en informatique musicale*, E. Marc Chemillier, Eds. Hermes, 1998, pp. 287-304.

[Castellengo 94] M. Castellengo, "La perception auditive des sons musicaux," in *Psychologie de la musique*, Zenatti, Ed. Paris: Puf, 94, pp. 55-87.

[Cooper, et al. 1952] F. Cooper, Delattre, Liberman, Borst and Gerstman, "Some Experiments on the Perception of Synthetic Speech Sounds," *J. of Acoust. Soc. Am.*, vol. 24, pp. 597-606, 1952.

[Freed 1990] D. Freed, "Auditory correlates of perceived mallet hardness for a set of recorded percussive sound events," *J. of Acoust. Soc. Am.*, vol. 87(1), pp. 311-322, 1990.

[Florens and Cadoz 1990] J.-L. Florens and C. Cadoz, *"Modèles et simulations en temps réel de cordes frottées,"* presented at 1er Congrès Français d'Acoustique, Lyon, 1990.

[Fourcade and Cadoz 1996] P. Fourcade and C. Cadoz, *"Sound Synthesis by Physical Modelling : an Elementary Striker,"* presented at Forum Acusticum, Antwergen, Conférence internationale avec comité de lecture, 96, vol. 82, p. Additional proceedings.

[Gibson 1966] J. J. Gibson, *The senses considered as perceptual systems*. Boston: Houghton Mifflin, 1966.

[Incerti 1996] E. Incerti, *"Synthèse de sons par modélisation physique de structures vibrantes : application pour la création musicale par ordinateur,"* Thèse de docteur ingénieur spécialité informatique, Grenoble

[Incerti 1997] E. Incerti, *"Modeling Methods for Sound Synthesis. Network Combinations and Complex Models for Physical Modeling: Application to Modes Clustering,"* presented at ICMC 97, Thessaloniki (Greece), 1997.

[Lindsay and Norman 1977] P. H. Lindsay and D. A. Norman, *Human information processing: An introduction to psychology*, 2nd édition ed. New York: Academic Press, 1977.

[Michaels and Carello 1981] C. F. Michaels and C. Carello, *Direct perception*. NJ: Prentice-Hall, 1981.

[Mangiarotti 1997] S. Mangiarotti, "Etude de l'attaque d'un son non-entrenu à partir d'excitateur multi-percussionnels," ACROE Grenoble, Rapport de stage, 1997.

[Risset 1994] J.-C. Risset, *"Modèles physiques et perception - Modèles physiques et composition,"* presented at Colloque Modèles physiques pour la création, Grenoble, 1994.

[Rosch 1978] Rosch, "Principles of Categorisation," in *Cognition and Categorization*, E. Rosch and B. B. Llyod, Eds. Hillsdale: Erbaum, L., 1978, pp. 27-47.

[Schaeffner 1968a] A. Schaeffner, "Chapitre premier : Origines corporelles," in *Origine des instruments de musique, Introduction ethnologique à l'histoire de la musique instrumentale*, deuxième édition, 1994 ed. Paris: Ecole des hautes-études en sciences sociales, 1968, p. 13.

[Schaeffner 1968b] A. Schaeffner, "Classification des instruments de musique," in *Origine des instruments de musique, Introduction ethnologique à l'histoire de la musique instrumentale*, deuxième édition, 1994 ed. Paris: Ecole des hautes-études en sciences sociales, 1968, p. Appendice.